Making Tensor Factorizations Robust to non-Gaussian Noise

Eric C. Chi[1] and Tamara G. Kolda[2]

[1]Department of Statistics,
Rice University

[2]Sandia National Laboratories, Livermore

December 10, 2010

CANDECOMP/PARAFAC (CP) Tensor Factorization

World View

Data $=$ Systematic Variation $+$ non-Systematic Variation

This talk

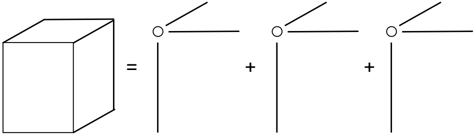
Systematic Variation: multilinear

Rank R approximation of $\mathcal{X} \in \mathbb{R}^{I \times J \times K}$.

$$\mathcal{X} = \mathcal{M} + \mathcal{E}$$

$$\mathcal{M} = \sum_{r=1}^{R} \mathbf{u}_r \circ \mathbf{v}_r \circ \mathbf{w}_r$$

$$m_{ijk} = \sum_{r=1}^{R} u_{ir} v_{jr} w_{kr}$$



Fitting the CP model

Minimize sum of transformed elementwise residuals

$$\min_{\mathbf{U},\mathbf{V},\mathbf{W}} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \rho(x_{ijk} - m_{ijk})$$

Minimize by block coordinate descent

Fix **V** and **W**.

$$\min_{\mathbf{U}} \sum_{i=1}^{I} \sum_{j=1}^{J} \sum_{k=1}^{K} \rho(x_{ijk} - m_{ijk})$$

Repeat fixing two factors and minimizing the other.

	$\rho(y) = y^2$	$\rho(y) = \lvert y \rvert$
MLE if e_{ijk}	i.i.d. Gaussian	i.i.d. Laplacian
Algorithm	CPALS	CPAL1

Violating Gaussian assumptions: Who cares?

- What kind of non-Gaussianity is problematic?
    - Sparse large perturbations.
- Prior work: matrices
    - Hawkins, Liu, and Young (2001)
    - Ke and Kanade (2005)
    - Zhou, Li, Wright, Candès, and Ma (2010)
- Prior work: tensor
    - Vorobyov, Rong, Sidiropoulos, and Gershman (2005)
        - Minimize 1-norm loss with block coordinate descent + linear programming

Majorization-Minimization

Strategy

Minimize a surrogate function that **majorizes** the objective.

Choose surrogate such that

- $\downarrow$ surrogate $\implies$ $\downarrow$ objective.
- surrogate is easier to minimize than objective.

Definition

Given f and g, real-valued functions on $\mathbb{R}^p$, g **majorizes** f at x if

1. $g(x) = f(x)$
2. $g(u) \geq f(u)$ for all u.

Majorizing an approximation

Smooth Approximation

$$\sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}|x_{ijk} - m_{ijk}| \approx \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sqrt{(x_{ijk} - m_{ijk})^2 + \epsilon},$$

for some small $\epsilon > 0$ ($\sim 1e$-10)
and $m_{ijk} = \sum_{r=1}^{R} u_{ir} v_{jr} w_{kr}$.

Block Coordinate Descent on approximate loss

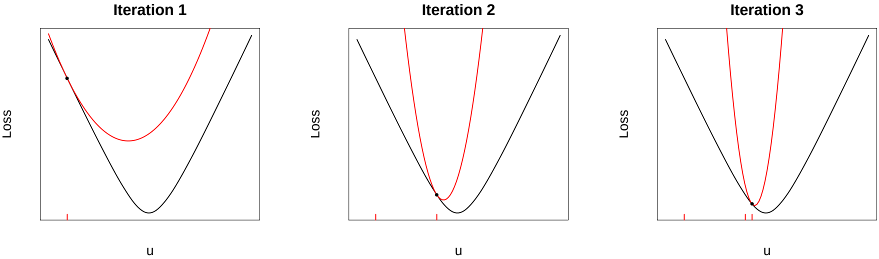
$$\min_{\mathbf{U}} \sum_{i=1}^{I}\sum_{j=1}^{J}\sum_{k=1}^{K}\sqrt{(x_{ijk} - m_{ijk})^2 + \epsilon}$$

- Problem separates in rows of $\mathbf{U}$.
- Each row, $\mathbf{u}_{(i)} \in \mathbb{R}^R$, can be fit with Iterative Reweighted Least Squares independently of all other rows.

MM Algorithm

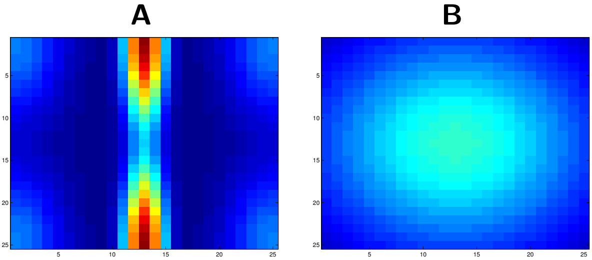
```
$g(\cdot|\mathbf{x}^{(0)}) \leftarrow$ majorization of $f$ at $\mathbf{x}^{(0)}$
repeat
    $\mathbf{x}^{(k+1)} \leftarrow \text{argmin}_{\mathbf{x}}\ g(\mathbf{x}|\mathbf{x}^{(k)})$
    $g(\cdot|\mathbf{x}_{k+1}) \leftarrow$ majorization of $f$ at $\mathbf{x}^{(k+1)}$
until convergence
```

$$\text{Loss} = \sum_i \sqrt{(x_i - u)^2 + \epsilon}$$

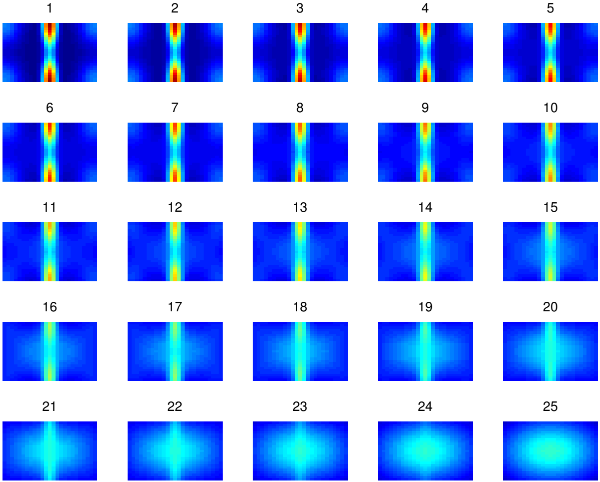


Toy example

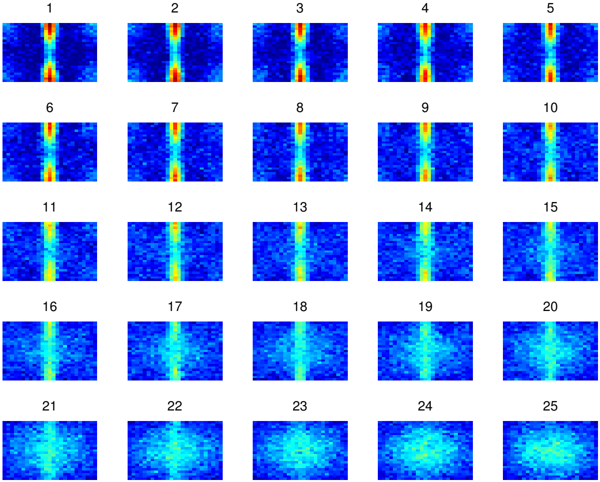
- $\mathcal{X} \in \mathbb{R}^{25 \times 25 \times 25}$.
- Slice $=$ mix of **A** and **B**.
- $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{25 \times 25}$.
- True rank $R = 2$.



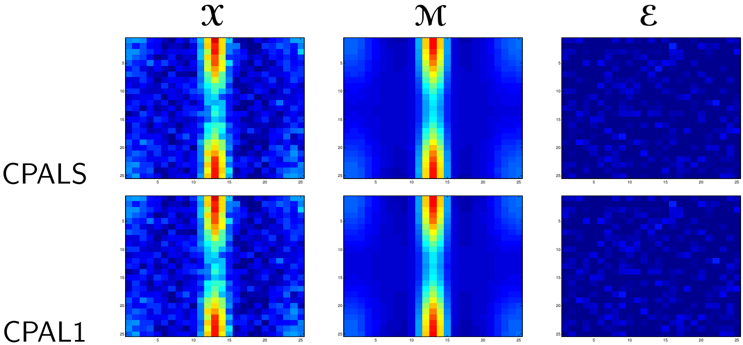
Toy example



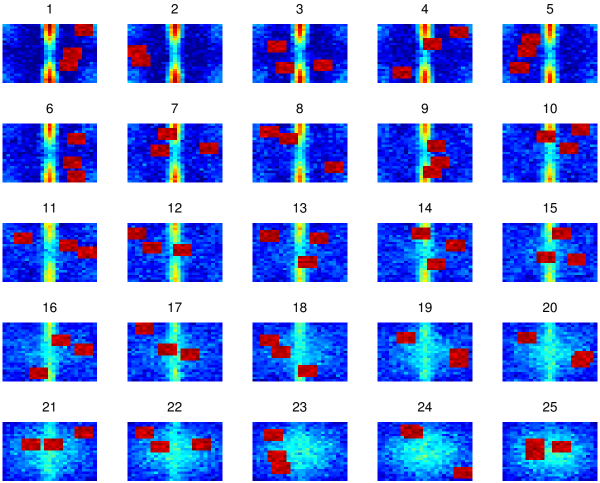
Toy example: Gaussian noise



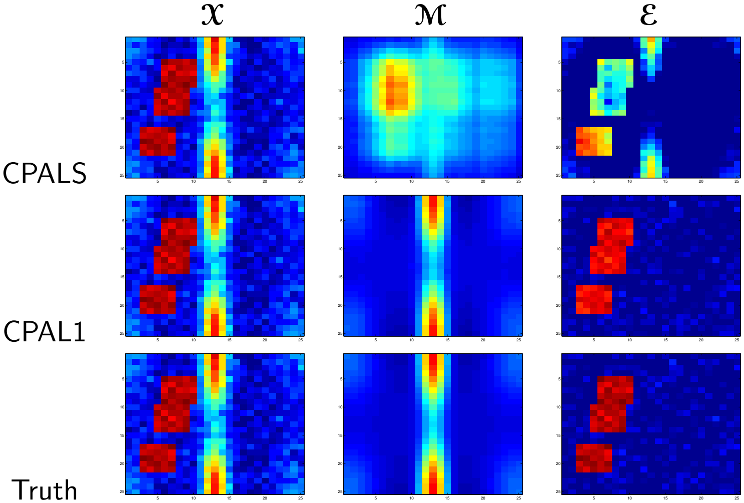
Gaussian Noise



Gaussian + non-Gaussian noise



Gaussian + non-Gaussian Noise



Discussion

Costs

- Computational: 1-norm minimization is more work than least squares.
- Statistical: Robustness versus efficiency tradeoff

Take home lesson

- Least squares can be sensitive to non-Gaussian perturbations.
- MM algorithms
  - Practical
  - Existing results on convergence
  - Existing methods for speeding up convergence
  - Majorizing losses other than 1-norm

Discussion

Future work

- Better robust loss functions?
- Data on different scales:
  - Binary
  - Non-negative data.

References for this work

- Extended abstract on arXiv
- Technical Report, in preparation
- Matlab code to be available online.

Eric C. Chi
echi@rice.edu